A gene refers to the sequence of DNA that encodes a single protein. Proteins are the molecular machines that carry out all the tasks performed by and in living cells (e.g. growth, digestion, motion, and photosynthesis). All proteins are made by cell machinery that transcribes DNA to mRNA and then translates mRNA into a chain of amino acids which fold into functioning 3-dimensional proteins. This transcription and translation machinery, as well as the machinery that breaks down proteins, is tightly regulated and controls the life of the cell. Proteins encoded by one gene up or down regulate other proteins (i.e. increase or decrease their production), forming a network that determines cell growth, health, and disease. Knowing and understanding this network can enable cell-based diagnostics (this network is abnormal, this person is sick), personalized treatment (this drug addresses that abnormality), or rational drug discovery (this network node is critical for this disease, find a drug that inhibits it).

So far, the identification of even a small portion of gene networks has been achieved by consensus over many PhD theses in multiple cellular biology labs. This is because the networks are complex (a single human cell has ~30,000 types of proteins) and acquiring, validating, processing, and understanding cell-data is a herculean task (even with current micro-array methods, the cell must be exploded to spill its contents to allow anaysis of mRNA and protein content, thus only one time point of data is available per cell). Yet in-vivo real-time gene/protein data is coming. Micro- and nano-scale systems are now being built to monitor living cells. At UMD, Shapiro, Smela, and Abshire have an ongoing effort to monitor cells on chip that will yield such data in the far term. Our goal in this ISR seed is to begin to develop mathematics so that when this cell data is available it can be effectively used.

We intend to develop algorithms for inferring gene (protein) networks from time-course data. Our approach is to view the entire network as a dynamical system, whose state comprises the levels of expression of each gene. We hypothesize that the autonomous time-evolution of such a system can be specified by a directed graph whose vertices represent genes, while directed edges signify causal dependence.

**Preliminary Results**

Our algorithm development is based on directed information. Given time-course data $X_1(t)$, $X_2(t)$, ..., $X_N(t)$, for mRNA or protein expression as a function of time, directed information allows us to determine which genes causally influence gene $K$. These are the genes that connect to gene $K$, and the edges between them are the directed graph edges. Identifying all such edges reveals the cell gene network.

We have used the concept of directed mutual information (DMI) as a guiding principle for designing algorithms for edge detection. The basic idea is that if $X$ is a random variable, and $Y$ is a function of $X$ plus independent noise, then it holds that $DMI(X \rightarrow Y) > 0$ and $DMI(Y \rightarrow X) = 0$, where $DMI$ stands
for directed mutual information (first proposed in [Marko73]). It can be shown that directed mutual information, applied to a pair of stochastic processes, quantifies not only the information flow between them (in bits per time unit) but also the directionality [Marko73].

In our preliminary work [Mathai07], we derived and tested algorithms and statistics based on the concept of directed mutual information. We obtained good results, as illustrated by our simulations. (In the simulations, we took a known gene network. Simulated it. Added noise. And then inferred the gene network from this synthetic, noisy data.) An important advantage of DMI in the detection of directed edges is that its use does not require testing all possible functions, i.e., in the absence of noise \( \text{DMI}(X \to Y) > 0 \) and \( \text{DMI}(Y \to X) = 0 \) hold if and only if \( Y \) is some function of \( X \). This DMI criteria identifies an edge only if gene \( Y \) depends on gene \( X \), else there is no edge. This formulation is also independent of the alphabet of the underlying stochastic processes, although larger alphabets increase the computational complexity of the algorithms used for computing the sample DMI.

**Proposed work**

Our approach in [Mathai07] unveiled the following problems and challenges that we will now address:

- Our statistics rely on the computation of the sample entropy rate of a stochastic process. While it has been shown that the sample entropy is a suitable statistics because it is robust to noise, its computation complexity grows exponentially with the dimension of the memory of the stochastic process. As such, an important challenge is how to efficiently compute the sample DMI directly or how to obtain approximate statistics that lead to computationally efficient algorithms.

- Our algorithms do not perform well in the presence of loops (i.e. when gene A regulates B which regulates C which then regulates A). Such loops occur frequently in living cells. We need to investigate how to detect edges in the presence of loops, even if that means having a non-unique solution. Many times, resulting ambiguity can then be eliminated by biologically inspired consistency rules [Alon06].

The work in [Mathai07] does not provide optimality guarantees. As such, we believe that an important step is to cast our problem as a solvable hypothesis testing problem in terms of statistics that share the “information flow” interpretation of the DMI.

We also believe that the problem of identifying gene networks is analogous to some of the research done in the networking community. Of particular interest is the problem of traffic identification or “stepping stone detection” in networks [He06]. This formulation is based on the realistic assumption that nodes in the graph have finite memory, which leads to a tractable paradigm and very efficient algorithms, under the assumption that the stochastic processes are Poisson-renewal. We would like to investigate how this framework could be modified to address our problem of gene edge detection.

**Conclusion**

We know that real-time data for living cells is coming. At UMD Shapiro is part of a group to create the technology to enable such measurements. But we have no idea what to do with the data once it is available – we don’t have the mathematics to process the data to infer the structure of living cells. This ISR seed effort is a first step towards this goal: we believe directed information is the right idea to infer which gene is influenced by which other genes since it is the only criteria that is independent of cell probability density functions (which are not known) and instead works directly with the time signals.

If we can make even partial advances in this area, if we can create methods that will infer a part of a cells network from measured data, the impact will be high. First, we will be able to increase the rate of molecular biology discovery, leading to faster understanding of the genetic basis of disease. Second, it will allow diagnosis from miniscule samples: as in remove living cells from a patient (from saliva, blood, or tissue), infer their network, and hence identify pathologies. And third, it will point to treatment and drug discovery: choose a treatment or design a drug that addresses the pathological proteins/ mRNA (i.e. critical nodes) in the network. We expect initial results will lead first to NSF funding (in control/networks programs) and then to NIH funding once clinically relevant networks can be identified.
References


